Not Too Much Life

Global Genome Initiative

Understanding and Preserving the Genomic Diversity of Life

Jonathan Coddington

Smithsonian Institution

Just one genome!

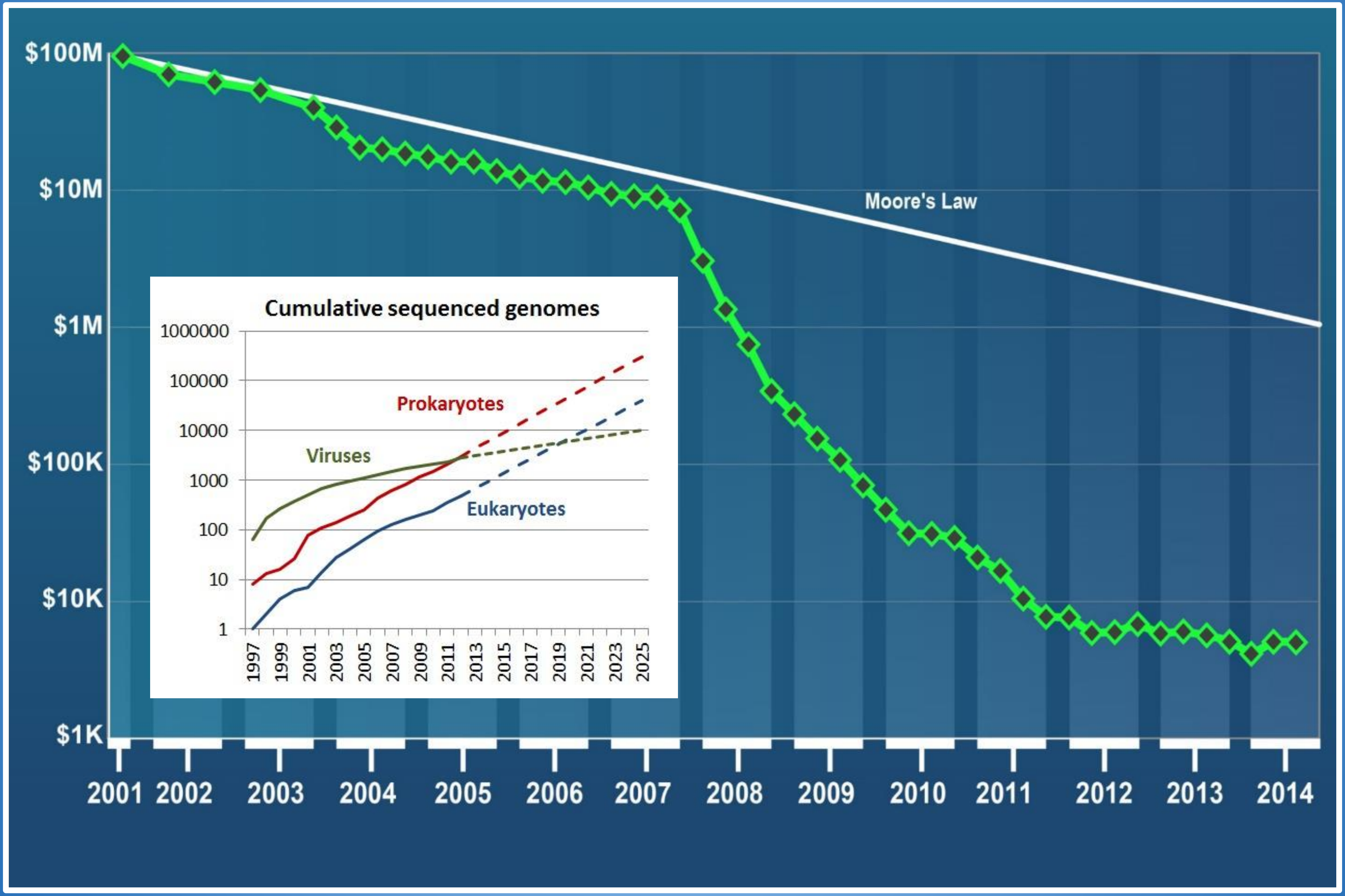
Outline

- Introduction & GGI Overview
- How big is Life?
- Feasibility
- *de novo* genomics (i5k) & new technologies

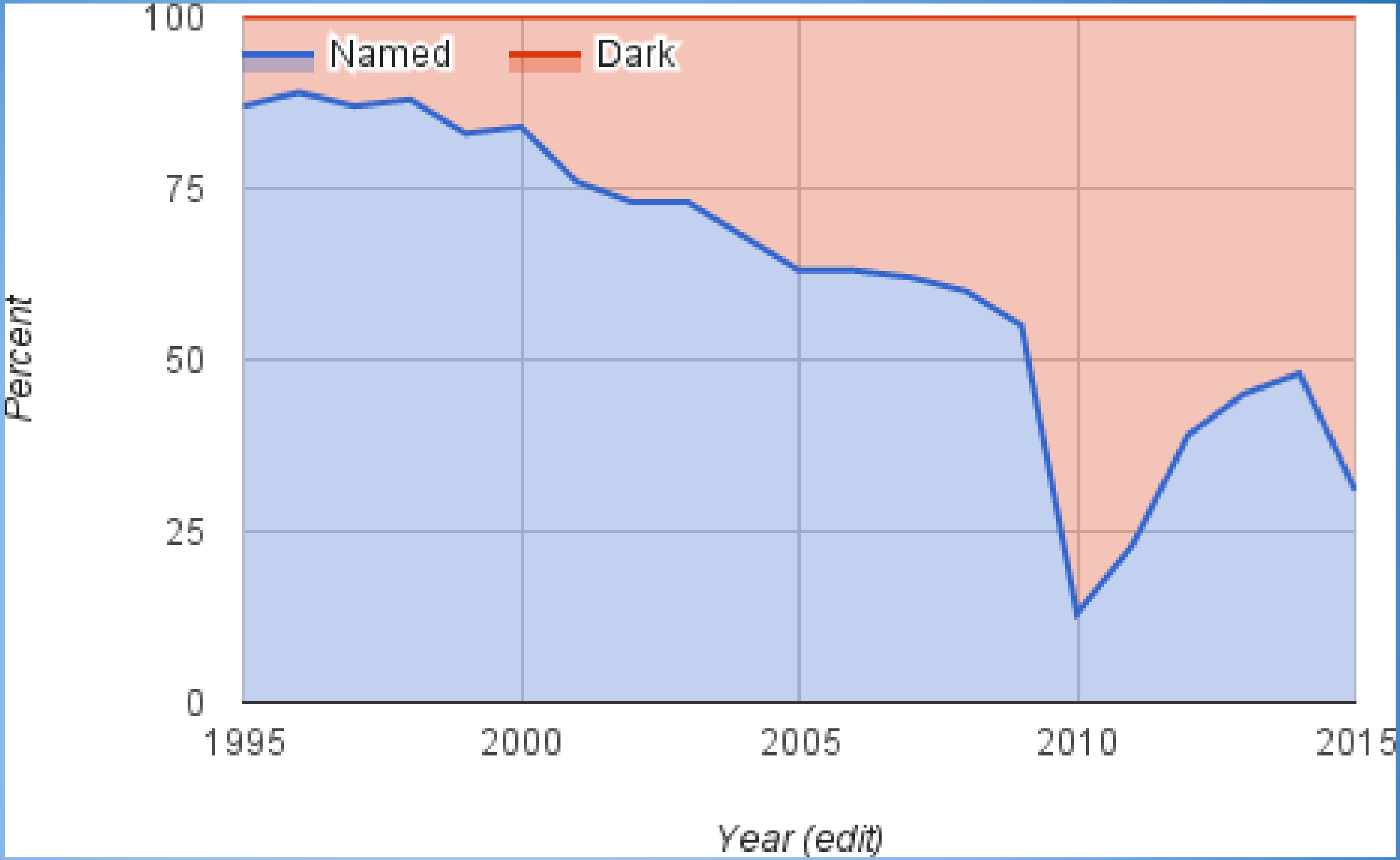
GGI Goals

Before	After
Sources: Hard-to-find, ambiguous quality tissues ambiguously owned by individual PI’s	Publically accessible, genome-quality tissues in enterprise biorepositories following best practices and Int. treaties
Data: Expensive “boutique” sequencing of a few model genomes	Affordable, coordinated, sequencing of a thoughtful synopsis of all of Life
Knowledge: Phenotype, expert-based taxonomy, underpinning environmental biology, evolution, conservation, ecology, biotech	Approximate taxonomic IDs of most organisms anywhere Cheap, precise, scalable tools

Biodiversity Genomics: costs and progress



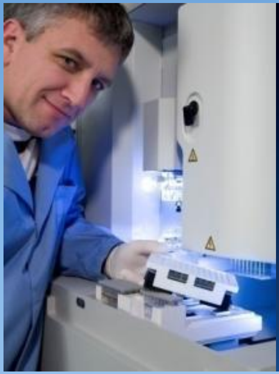
“Dark” Taxa

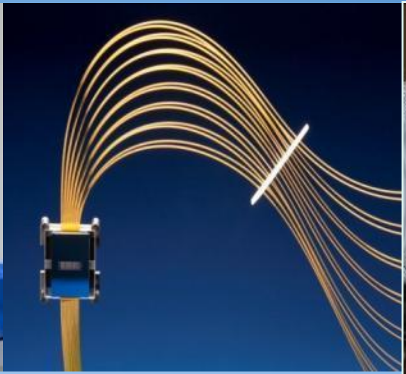


Dark taxa outpacing names (58% spiders)
Taxonomists dwindle
Practical, mesoscale ID’s urgently needed

R. Page 2016 PeerJ PrePrints 1649v1

Success (Phase I)





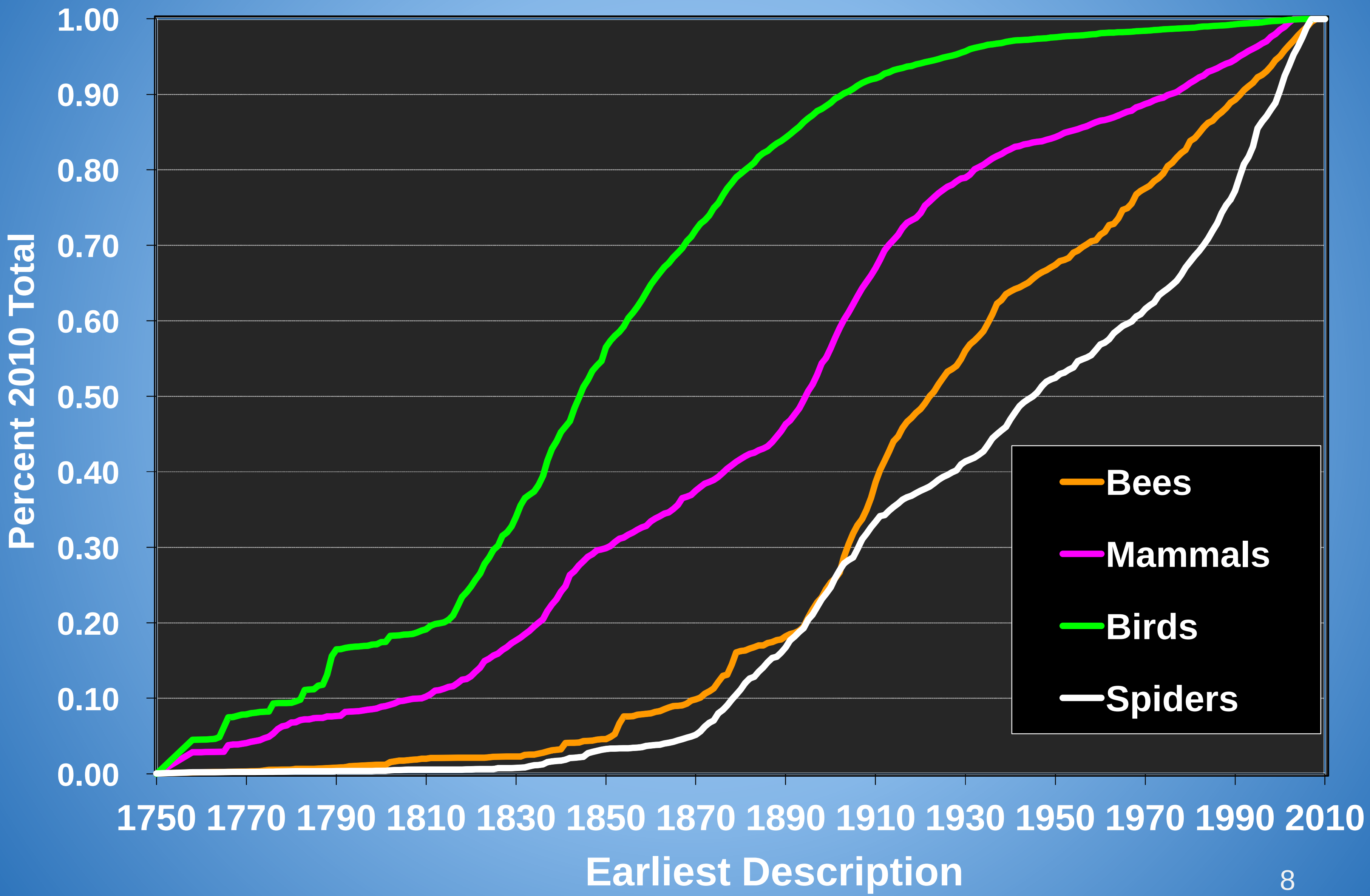
Preservation of 50 % of major branches of Life in 6 years (~10,000 families, 40-100,000 genera)

- Synergize research impact and productivity
- National and international partnerships and networks
- Approximate ID of any organism on Earth
- Global biorepositories & informatics
- Genome samples pivotal to new research outcomes
- Awareness and public understanding

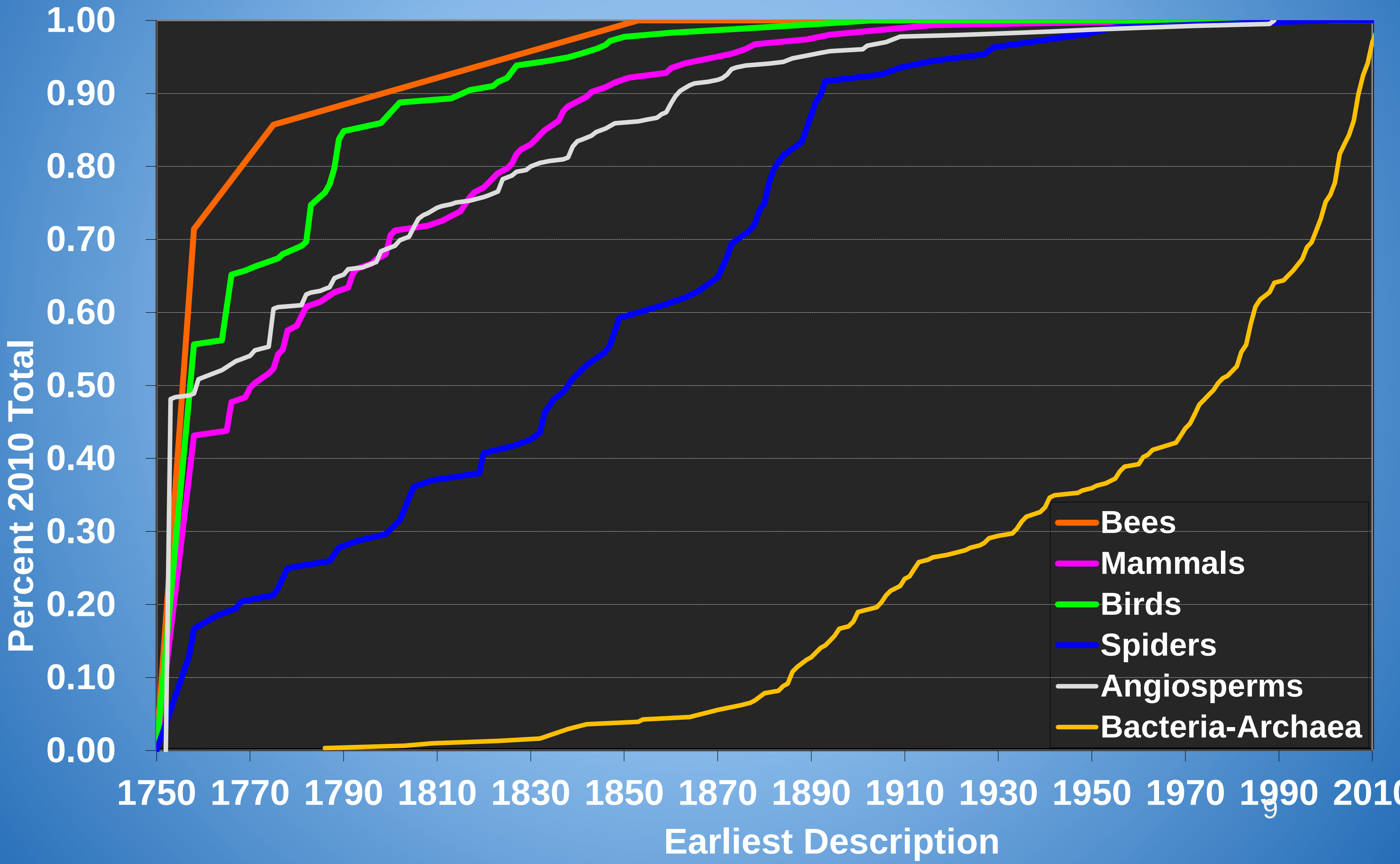
Outline

- Introduction & GGI Overview
- How big is Life?
- Feasibility
- *de novo* genomics (i5k) & new technologies

Discovery of “Species”

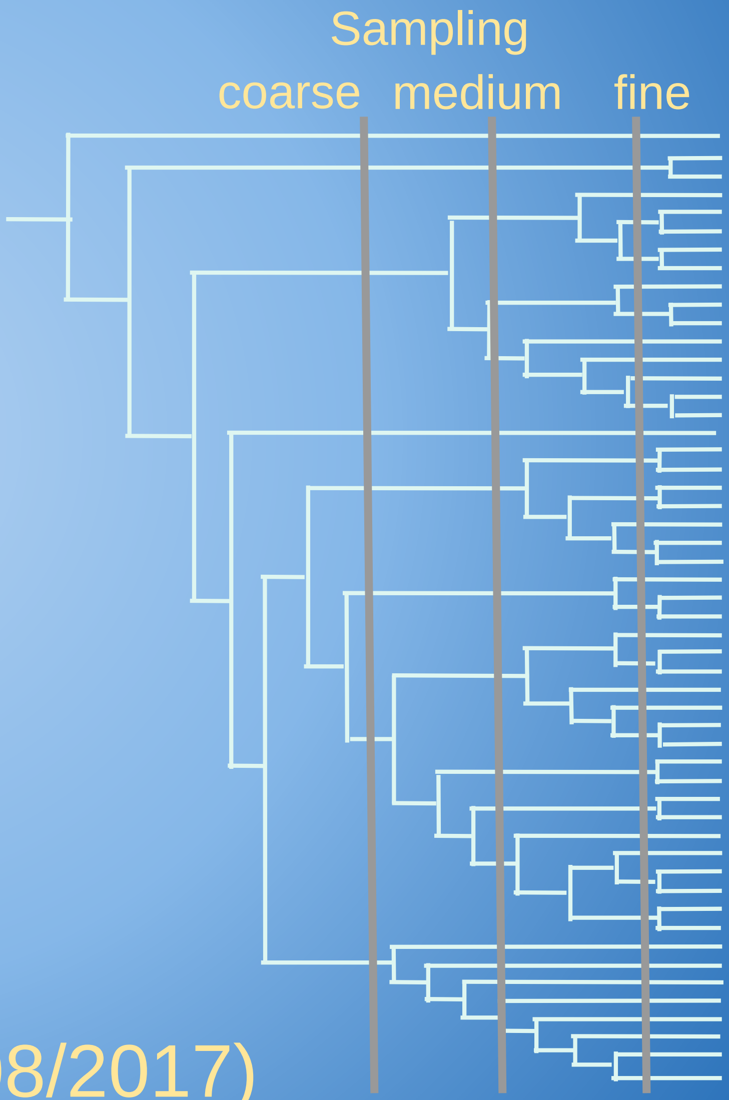


Discovery of “Families”



Feasibility: Phylogeny

Domains	3
Phyla/ Divisions	89
Classes	258
Orders	1,148
Families[1]	~8,850
Genera[2]	~149,000
Species	>15,000,000



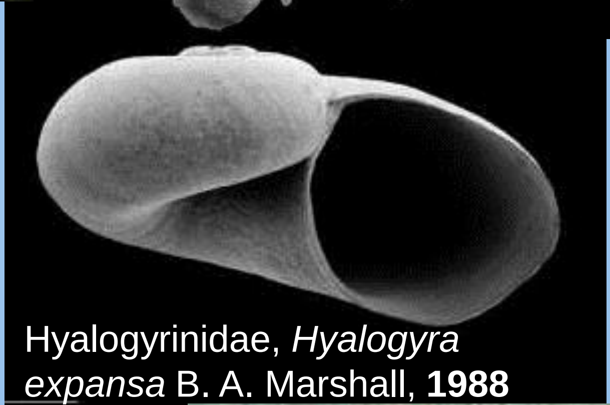
[1] 8,613, [2] 83,066 in Genbank (08/2017)

Gomortegaeae (*Gomortega keule* (Molina) Baill. **1972)**

Limnognathidae (*Limnognathia maerski* Kristensen & Funch **2000**)

Sapayoaidae (Sapayoa aenigma Hunt **1903**)

Godzilliidae (*Godzillius robustus* Schram *et al.*, **1986**)



Hyalogyrinidae, *Hyalogyra expansa* B. A. Marshall, **1988**

Trogloraptoridae (*Trogloraptor marchingtoni* Griswold, Audisio & Ledford, **2012**

Craseonycteridae (Craseonycteris thonglongyai. Hill, **1974**)

Protanguillidae **(***Protoanguilla palau* G. D. Johnson, H. Ida & Miya, **2012**)

Outline

- Introduction & GGI Overview
- How big is Life?
- Feasibility
- *de novo* genomics (i5k) & new technologies

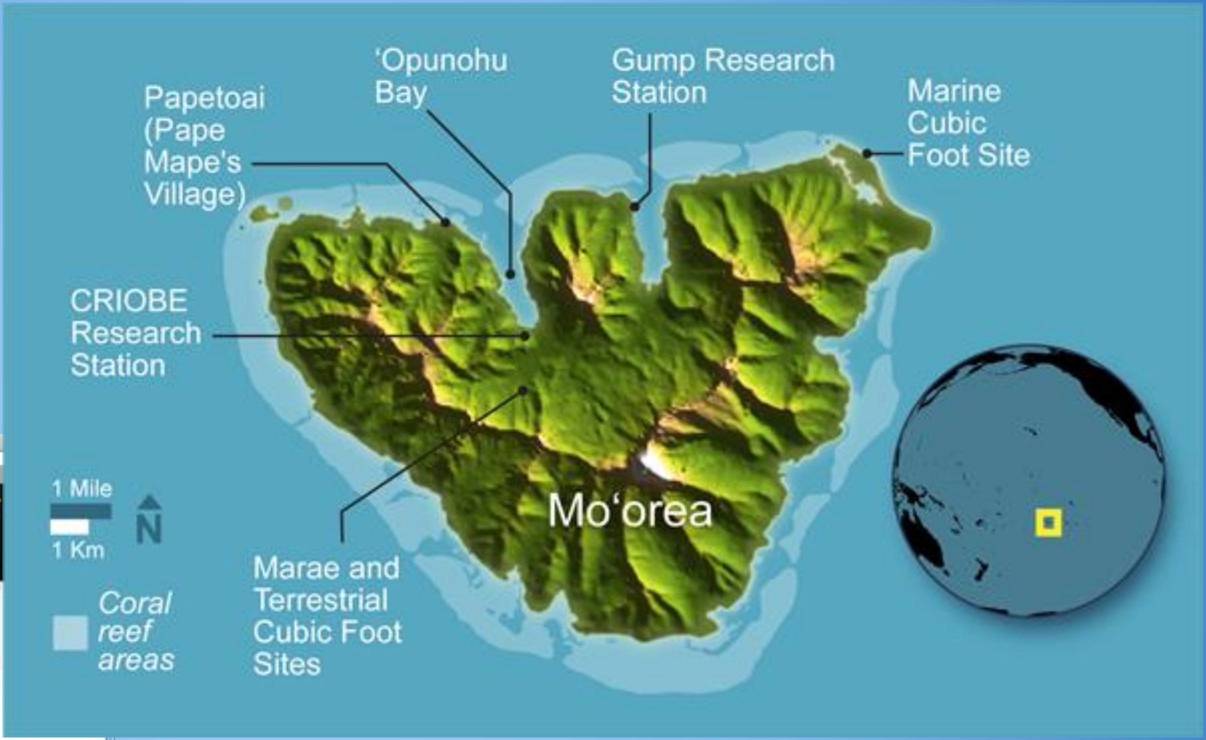
Feasibility: Moorea Biocode

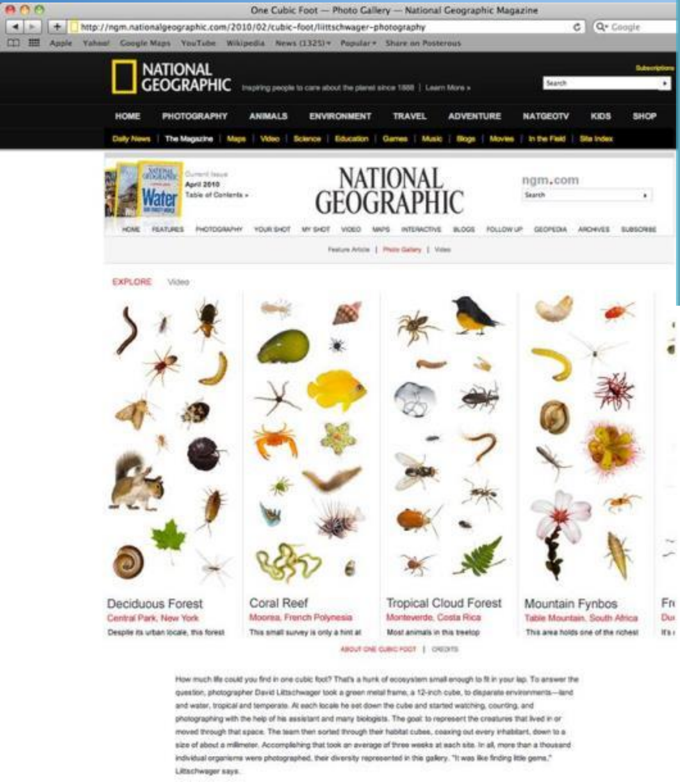
Phyla: 74%

Class: 61%

Order: 42%

Family: 23%




One Cubic Foot — Photo Gallery — National Geographic Magazine
http://ngm.nationalgeographic.com/2010/02/cubic-foot/liittschwager-photography
Apple Yahoo! Google Maps YouTube Wikipedia News (1325) Popular Share on Posterous
NATIONAL GEOGRAPHIC
Inspiring people to care about the planet since 1888 | Learn More »
Subscriptions
Search
HOME PHOTOGRAPHY ANIMALS ENVIRONMENT TRAVEL ADVENTURE NATGEOTV KIDS SHOP
Daily News | The Magazine | Maps | Video | Science | Education | Games | Music | Blogs | Movies | In the Field | Site Index
Current Issue
April 2010
Table of Contents »
Water
NATIONAL GEOGRAPHIC
ngm.com
Search
HOME FEATURES PHOTOGRAPHY YOUR SHOT MY SHOT VIDEO MAPS INTERACTIVE BLOGS FOLLOW UP GEOPEDIA ARCHIVES SUBSCRIBE
Feature Article | Photo Gallery | Video
EXPLORE Video
ABOUT ONE CUBIC FOOT | CREDITS

Deciduous Forest
Central Park, New York
Despite its urban locale, this forest

Coral Reef
Moorea, French Polynesia
This small survey is only a hint at

Tropical Cloud Forest
Monteverde, Costa Rica
Most animals in this treetop

Mountain Fynbos
Table Mountain, South Africa
This area holds one of the richest

How much life could you find in one cubic foot? That's a hunk of ecosystem small enough to fit in your lap. To answer the question, photographer David Liittschwager took a green metal frame, a 12-inch cube, to disparate environments—land and water, tropical and temperate. At each locale he set down the cube and started watching, counting, and photographing with the help of his assistant and many biologists. The goal: to represent the creatures that lived in or moved through that space. The team then sorted through their habitat cubes, coaxing out every inhabitant, down to a size of about a millimeter. Accomplishing that took an average of three weeks at each site. In all, more than a thousand individual organisms were photographed, their diversity represented in this gallery. "It was like finding little gems," Liittschwager says.

Japan
North
Pacific
Ocean
North
America
Hawaii
180°
165°
150°
135°
15°
EQUATOR
0°
Oceania
Australia
30°
New
Zealand
South
Pacific
Ocean
45°
Antarctica

Partner Institutions

Gordon and Betty MOORE FOUNDATION

Richard B. Gump South Pacific Research Station

Smithsonian Institution

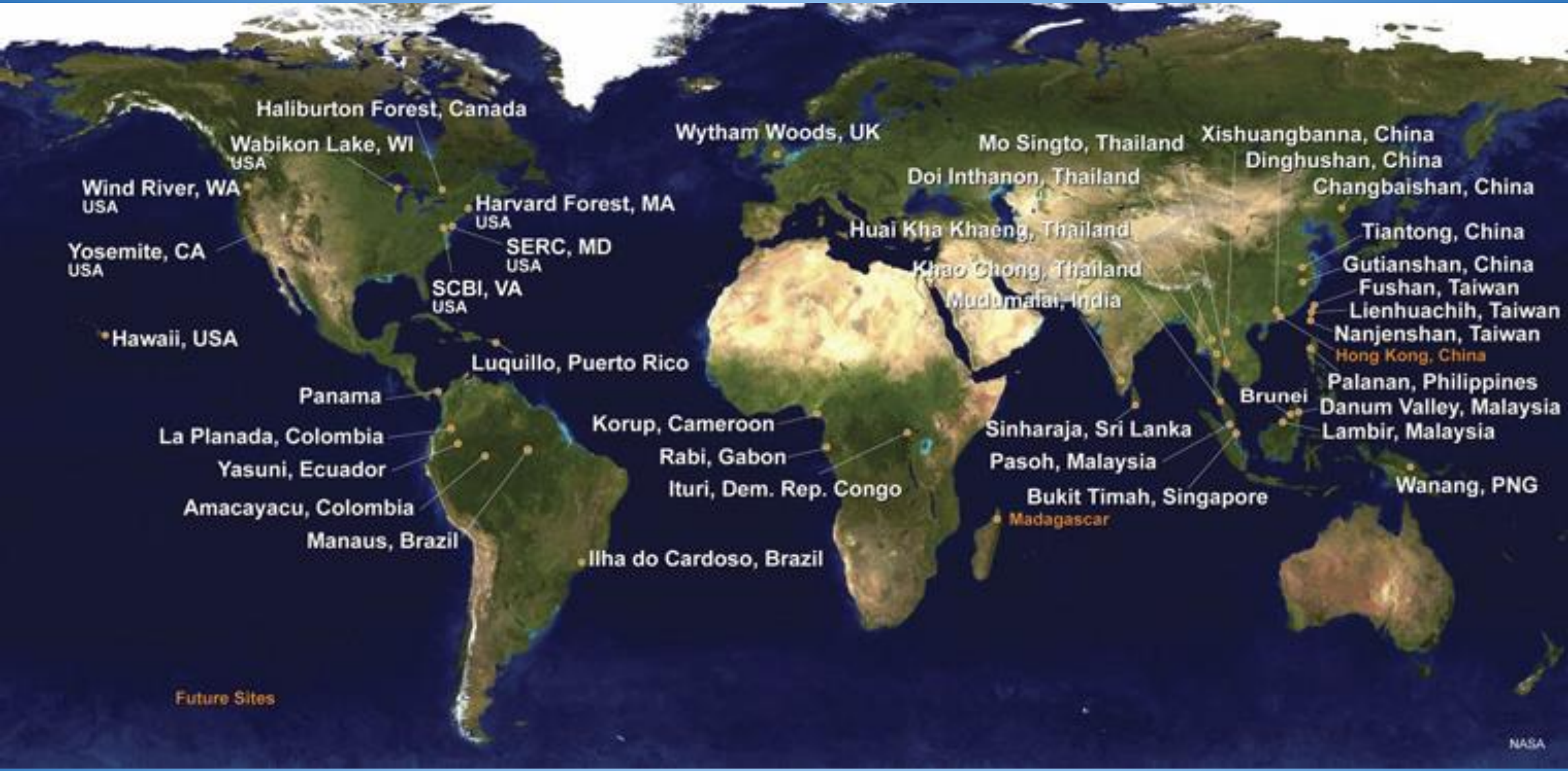
IRD Institut de recherche pour le développement

UF UNIVERSITY of FLORIDA

Bishop Museum

Ex worldatlas.com

Feasibility: Forest Global Earth Observatories



Smithsonian Institution Forest Earth Observatory

40 plots, 10,500 species, 4,346 genera (“trees”)

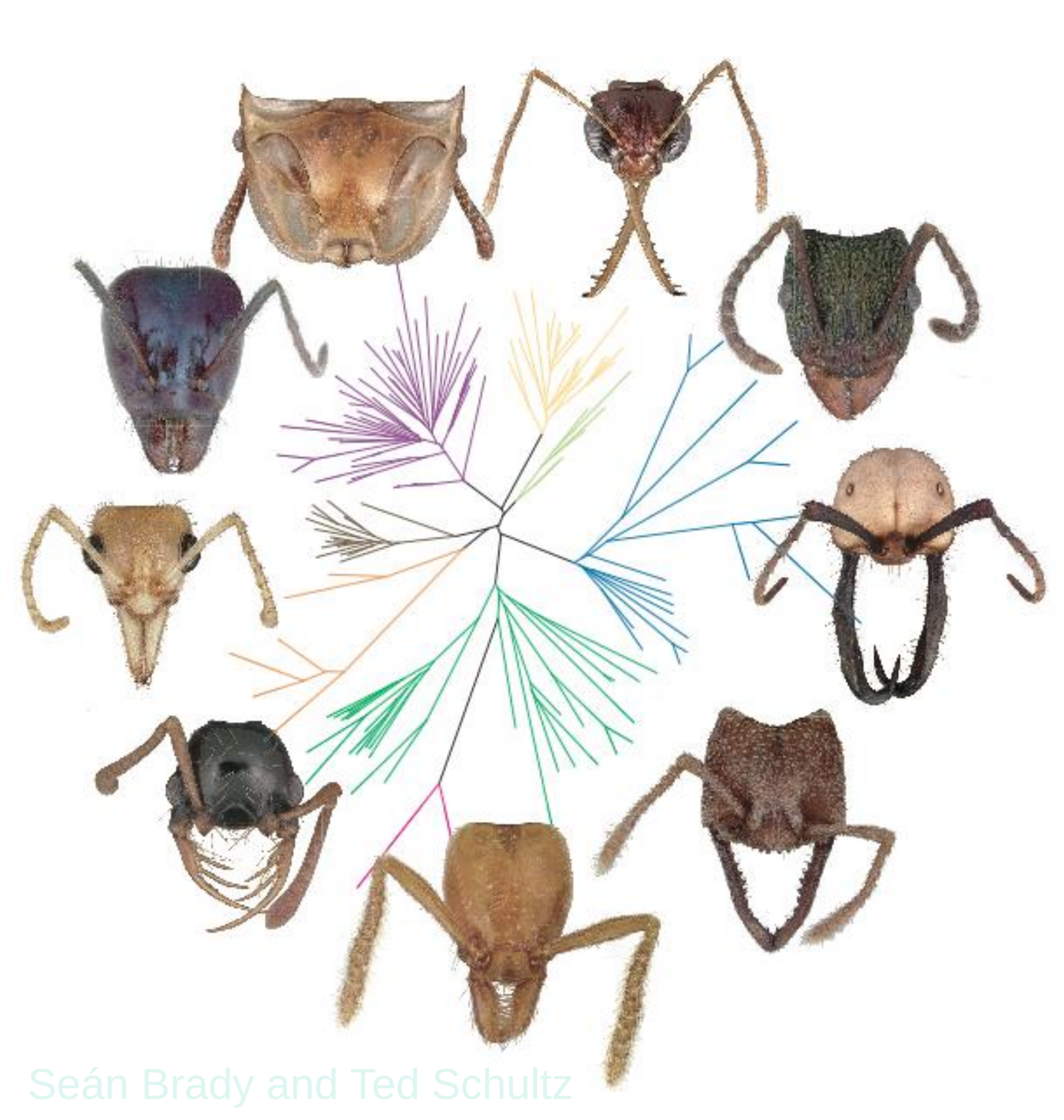
~60% world total?

Feasibility: Taxonomy

Ants

- 290 worldwide genera

- 240 genera w DNA extractions (82%)

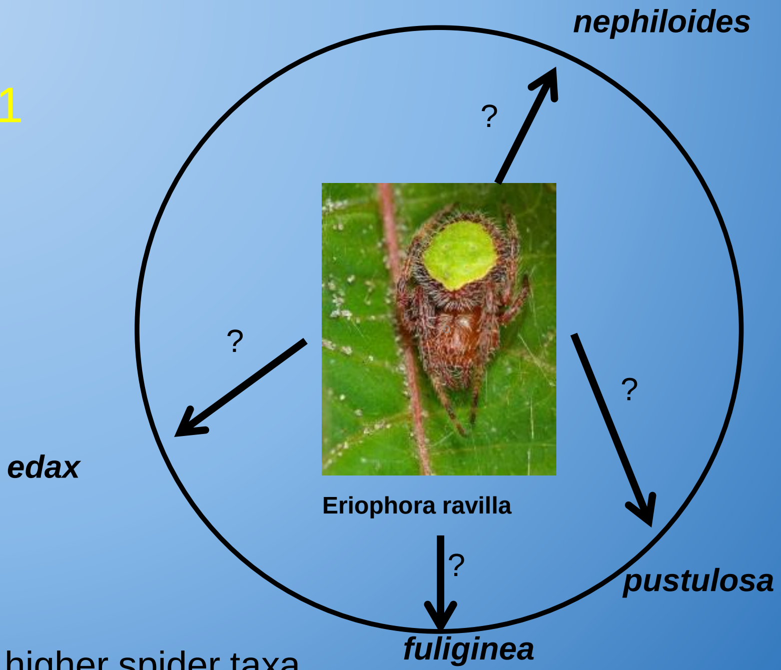


Seán Brady and Ted Schultz

Feasibility: Taxonomy

Barcode (COI) ID “radius” (e,g, European spiders)

- 50 families, 313 genera, and 821 species
- 873 sequences blasted against themselves
- 91% correct at family level[1]
- 85% correct at genus level[1]



[1]PIdent >0.95

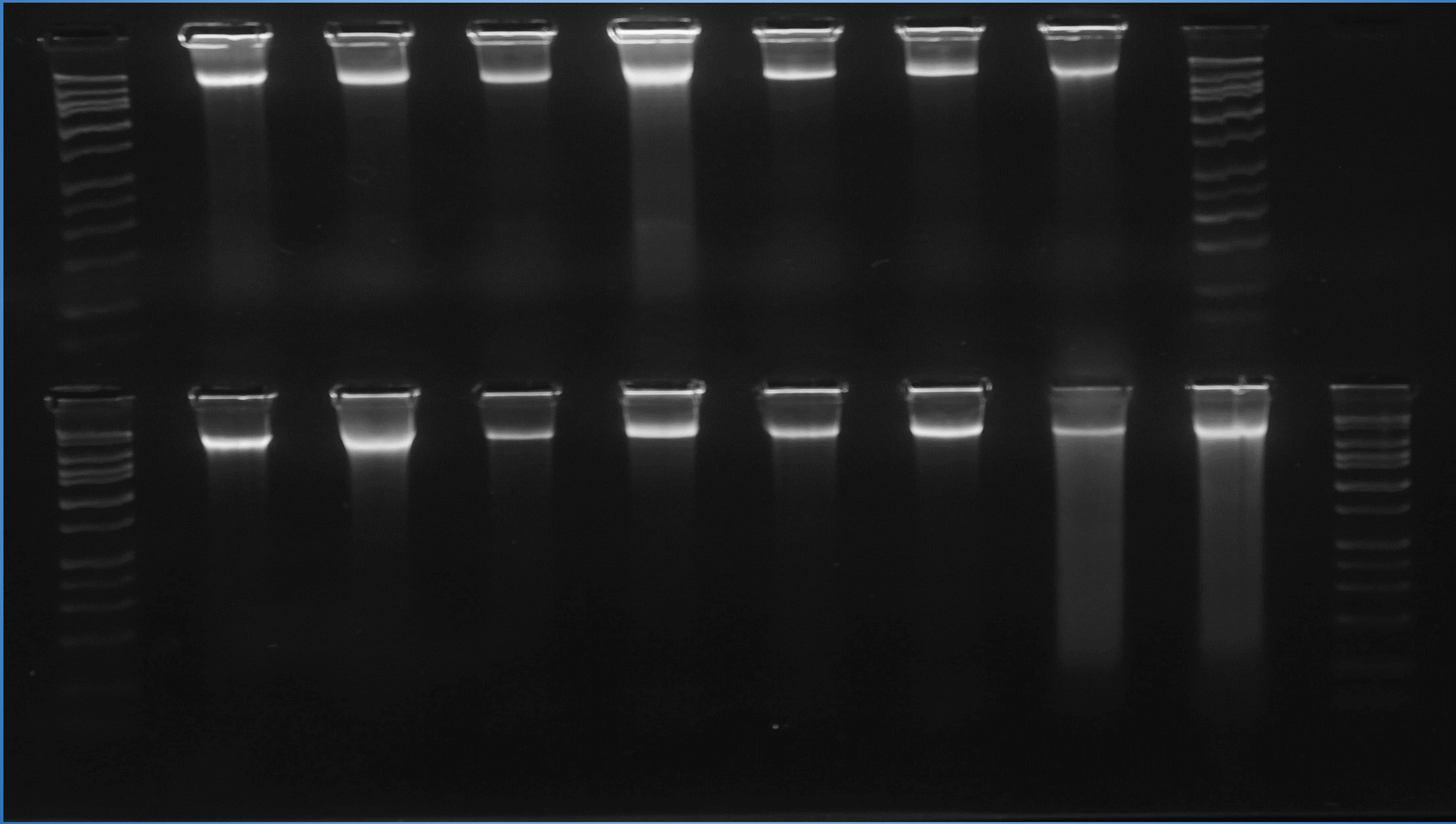
Coddington & Al. 2016. DNA barcode data accurately assign higher spider taxa

NMNH Biorepository 4-5M 2ml tube capacity

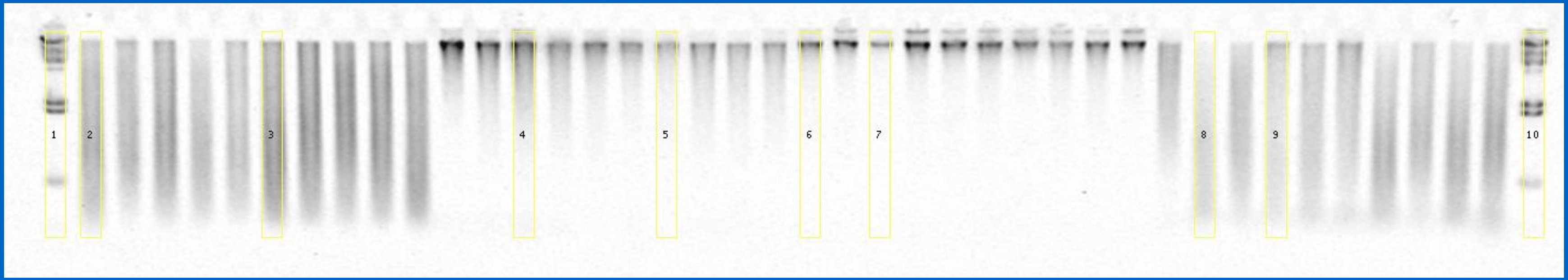
58 Freezers

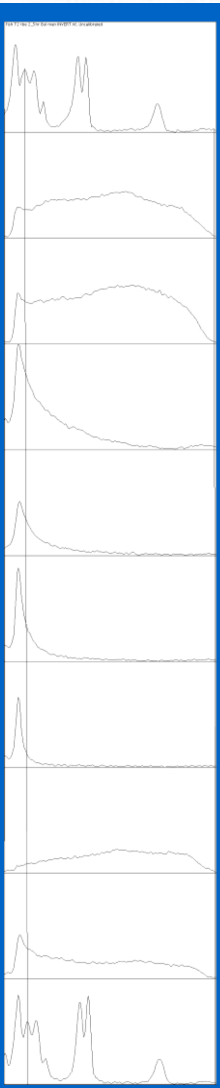
24 Nitrogen Tanks

High Molecular Weight DNA



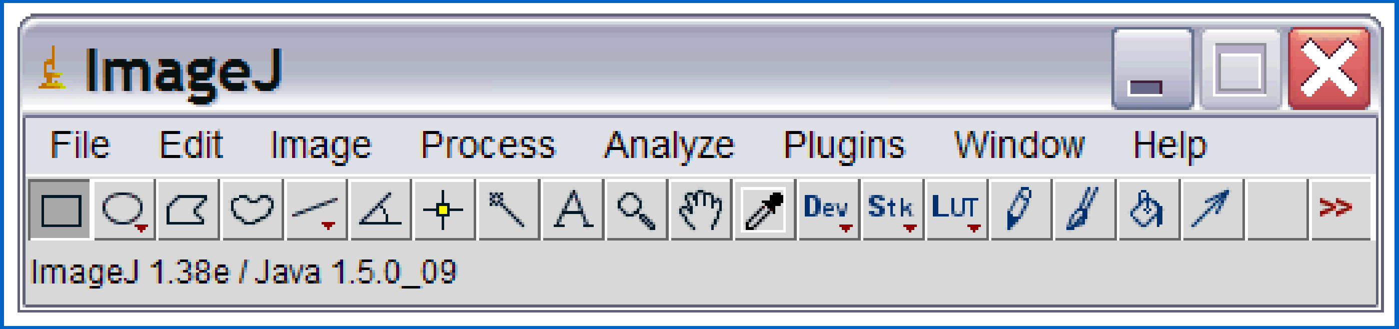
Genome Quality Tissues + Vol, Conc.





ImageJ

Image Processing and Analysis in Java

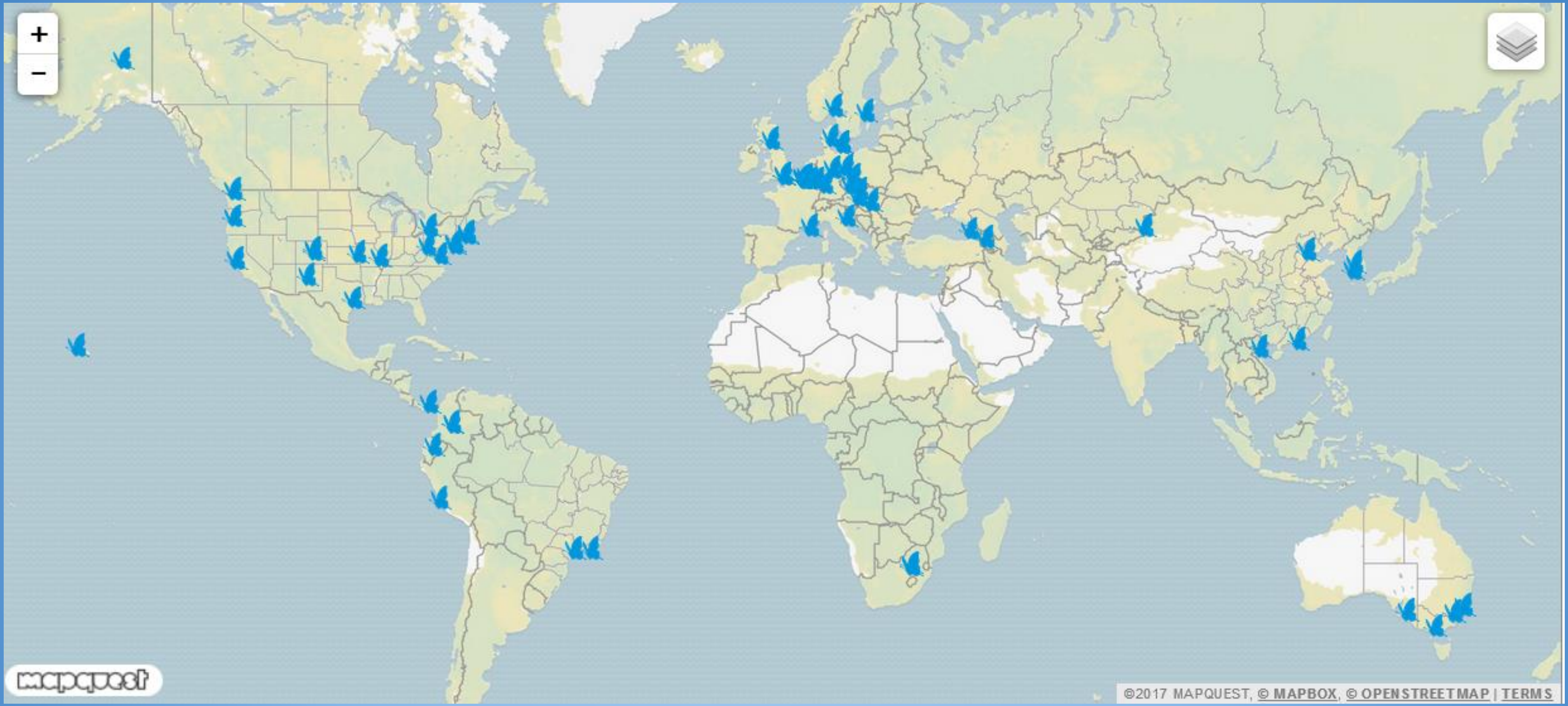


ImageJ 1.38e / Java 1.5.0_09

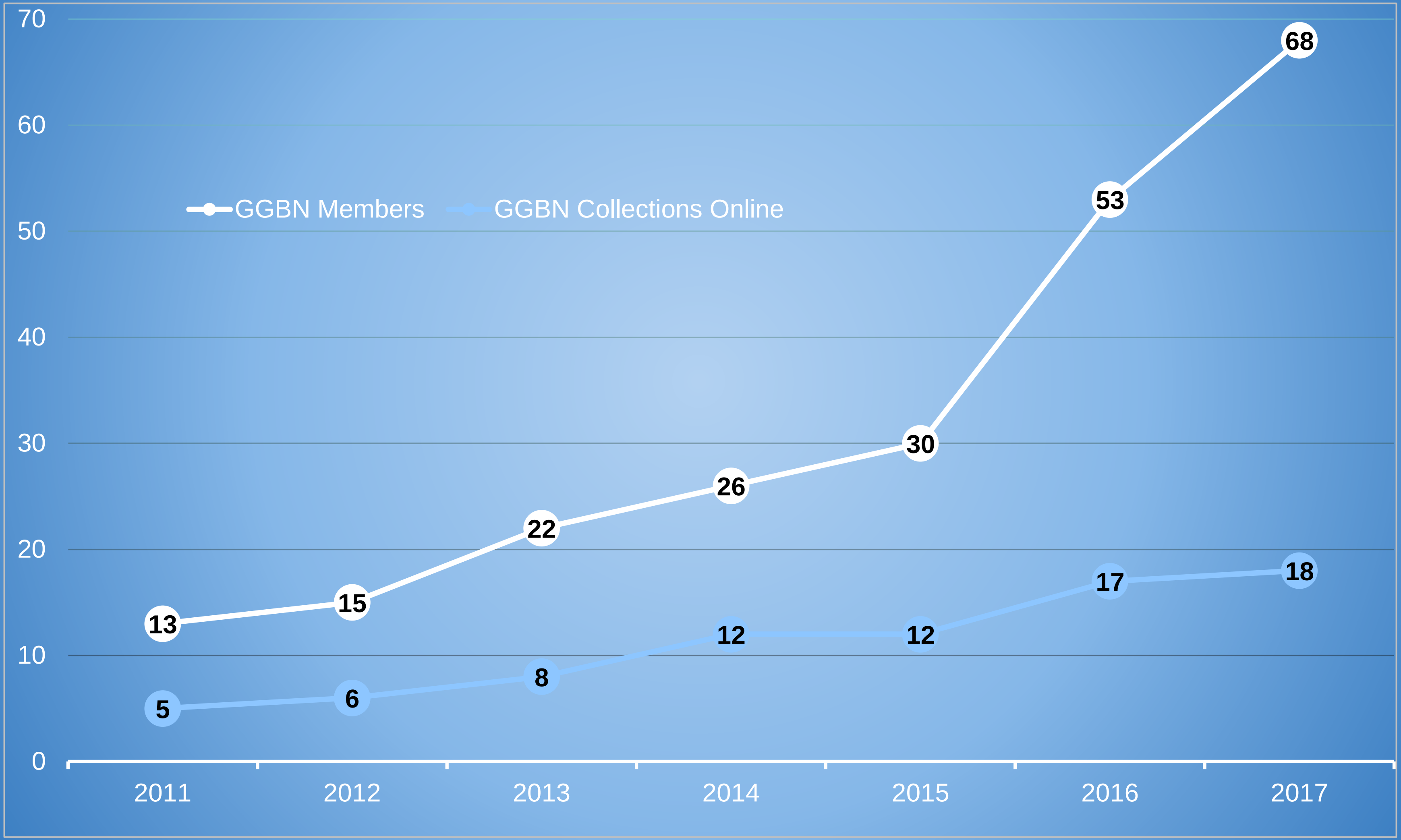
Column	> 9416 bp	< 9416 bp	total	% > 9416 bp
2	3503.234	61565.425	65070.659	5.4%
3	5026.648	79952.588	84982.236	5.9%
4	12537.255	27177.505	39718.76	31.6%
5	5965.548	8789.3	14759.848	40.4%
6	8906.134	6386.066	15298.2	58.2%
7	5767.962	2612.631	8387.593	68.8%
8	1096.042	29235.655	30339.697	3.6%
9	4876.841	28195.785	33081.626	14.7%

http://imagejdocu.tudor.lu/doku.php?id=video:analysis:gel_quantification_analysis

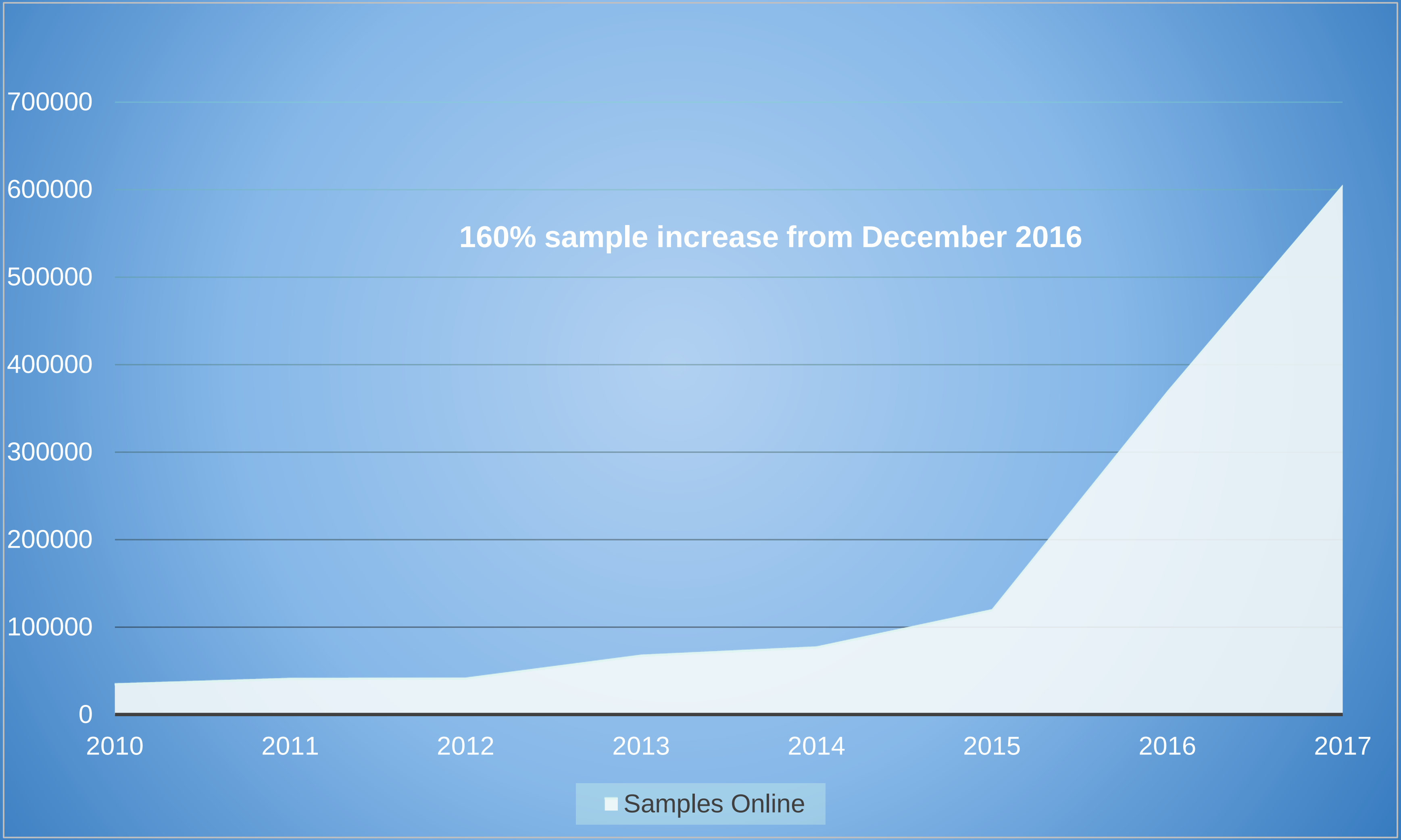
68 GGBN members, 22 countries, 603,902 samples, 2,824 families, 14,116 genera



GGBN Member and Collections Growth



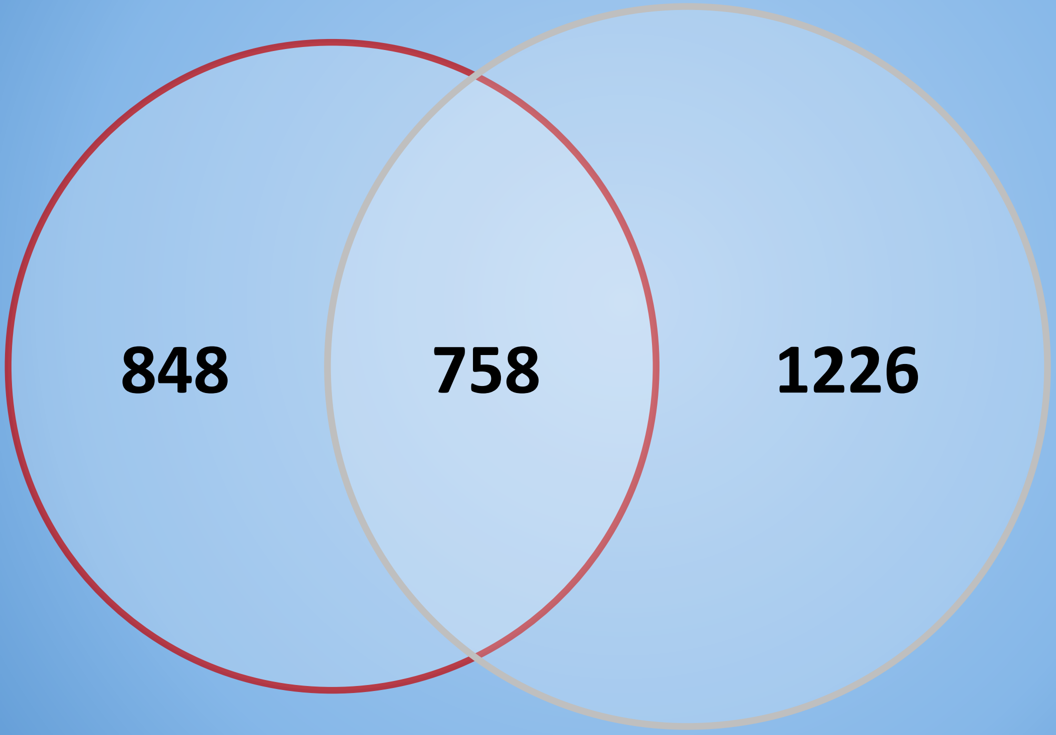
GGBN Online Sample Growth



160% sample increase from December 2016

GGBN Families

NMNH, Washington (1606)

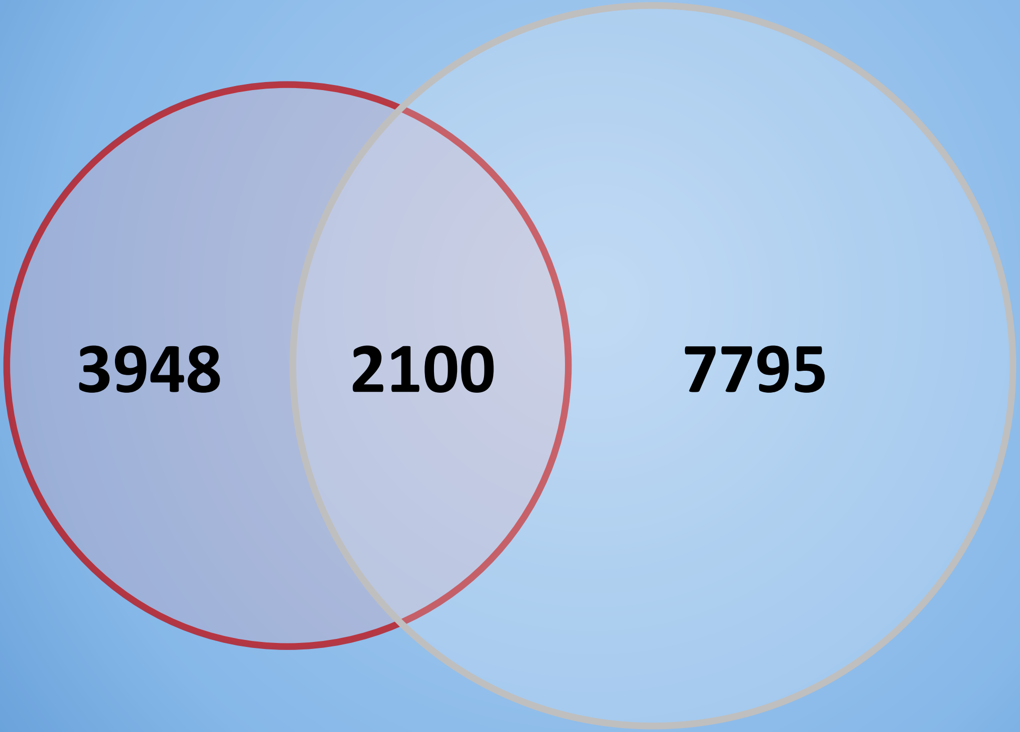


All Other Institutions (1984)

AIT, Tulln
BGBM, Berlin
CUni, Prague
DBG, Denver
DSMZ, Braunschweig
IRB, Rovinj
IVB, Brno
MfN, Berlin
NHM, London
NHMD, Copenhagen
NHMO, Oslo
NYBG, New York
OGL, Nahant
QCAZ, Quito
RBGK, London
Senckenberg, Frankfurt
ZFMK, Bonn

As of 24 Oct 2017

GGBN Genera



NMNH, Washington (6048)

All Other Institutions (9895)

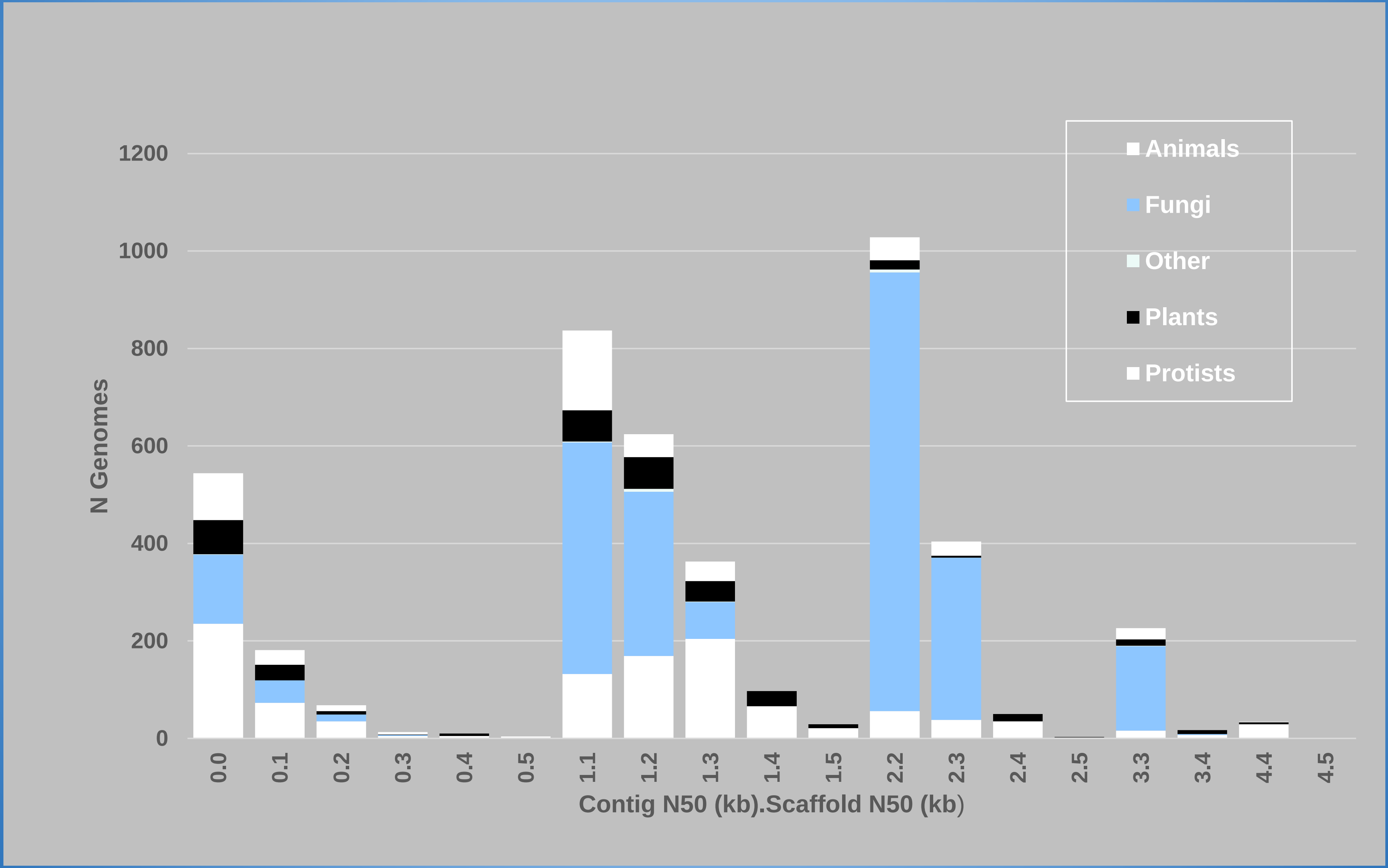
AIT, Tulln
BGBM, Berlin
CUni, Prague
DBG, Denver
DSMZ, Braunschweig
IRB, Rovinj
IVB, Brno
MfN, Berlin
NHM, London
NHMD, Copenhagen
NHMO, Oslo
NYBG, New York
OGL, Nahant
QCAZ, Quito
RBGK, London
Senckenberg, Frankfurt
ZFMK, Bonn

As of 24 Oct 2017

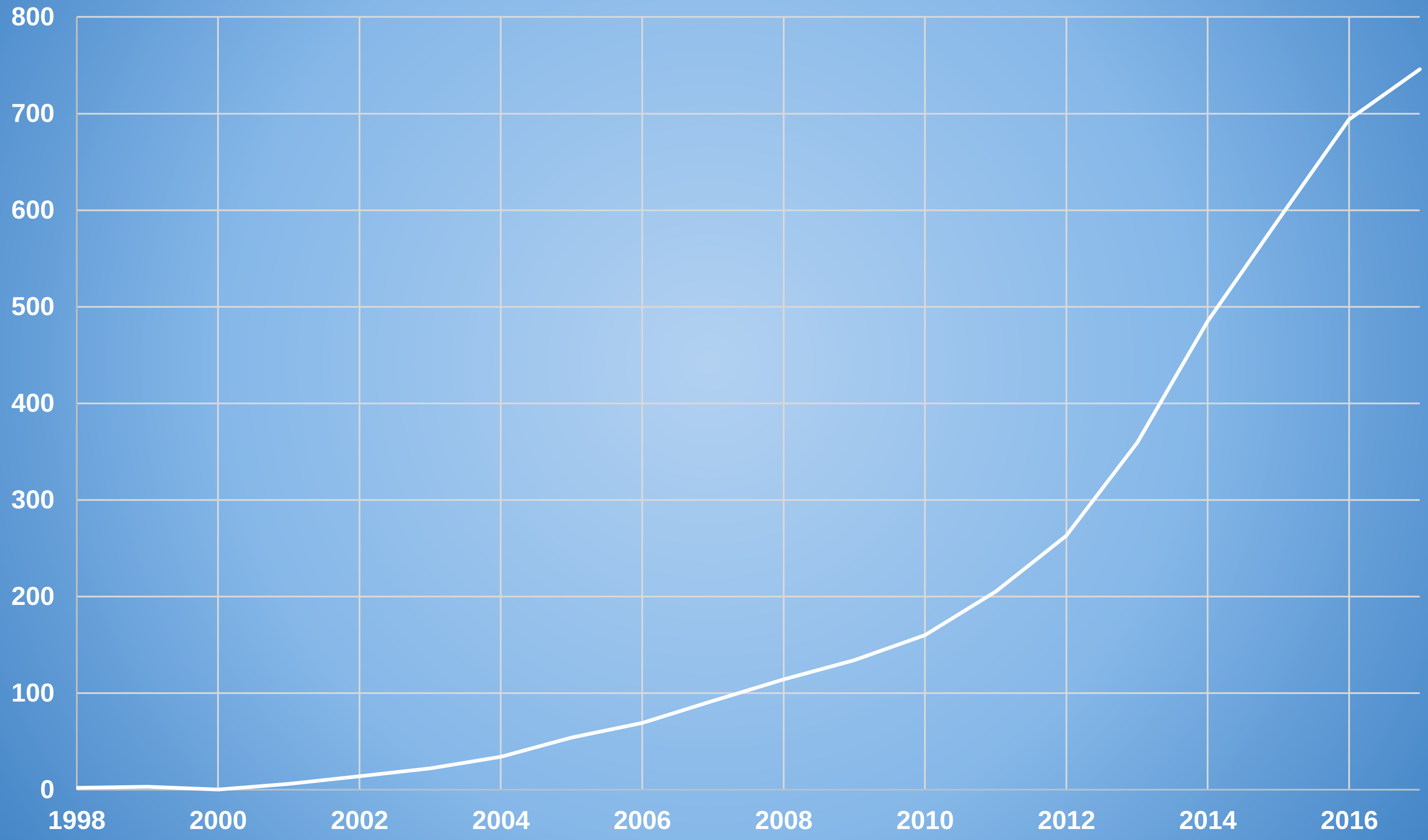
Outline

- Introduction & GGI Overview
- How big is Life?
- Feasibility
- *de novo* genomics (i5k) & new technologies

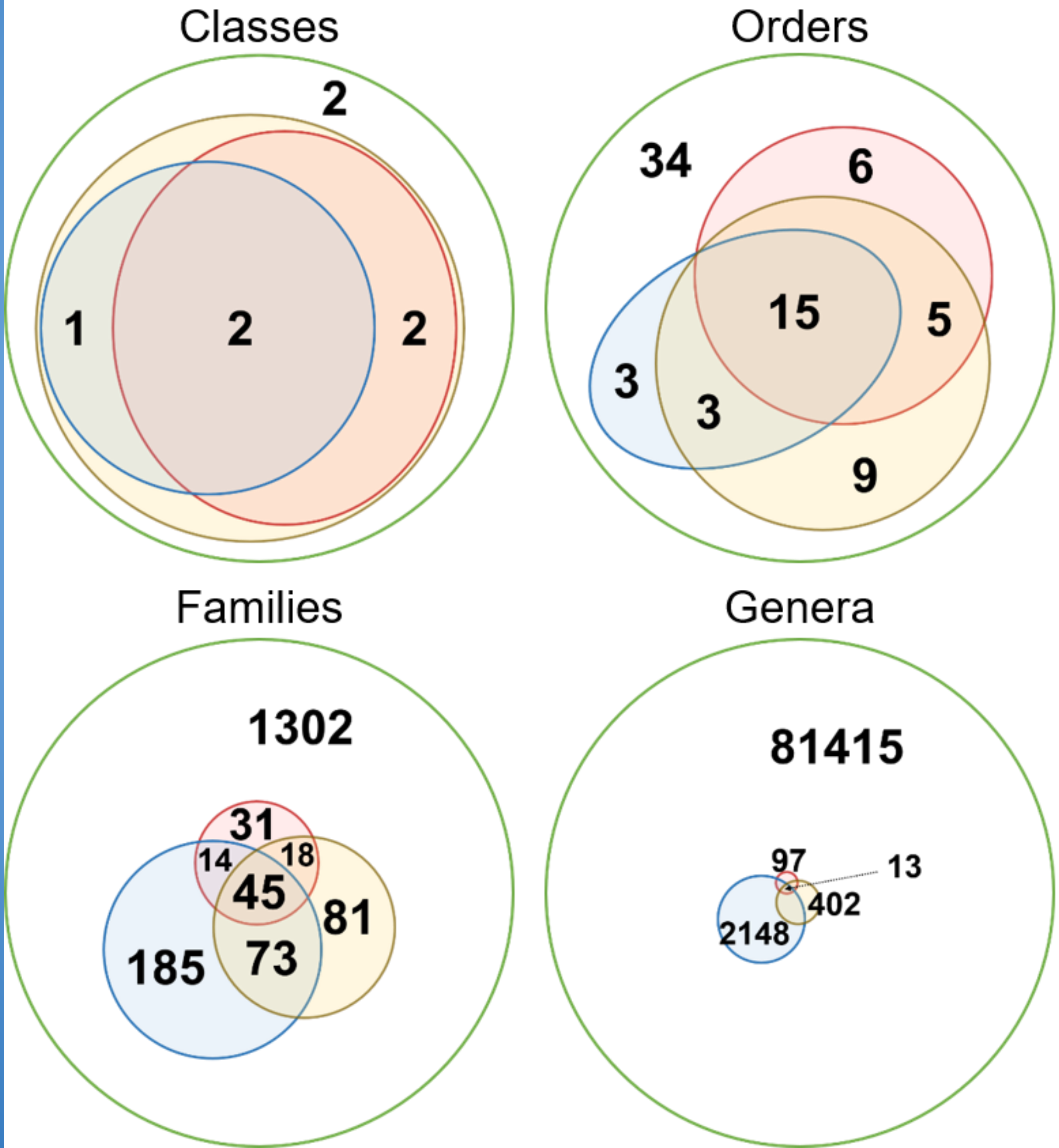
Eukaryotic Genome Quality (n=4,533)



Cumulative Count of Families w Genome



Terrestrial Arthropods



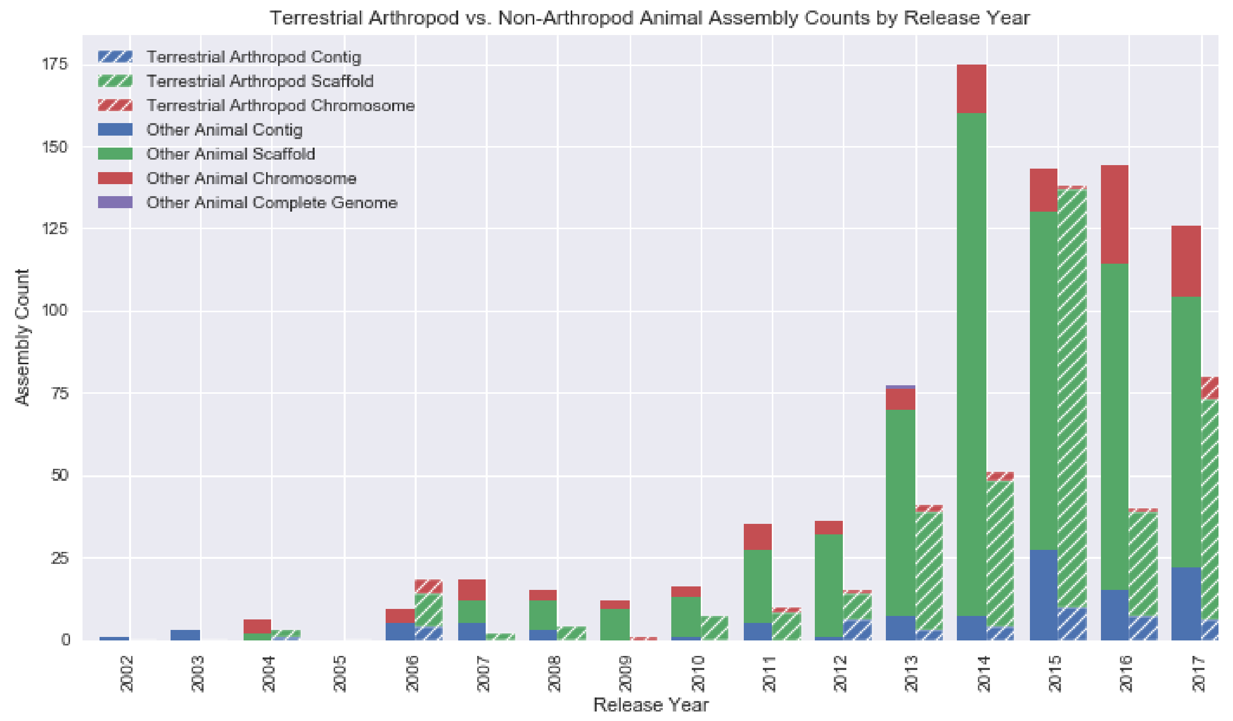
Red:
National Center for Biotechnology Information

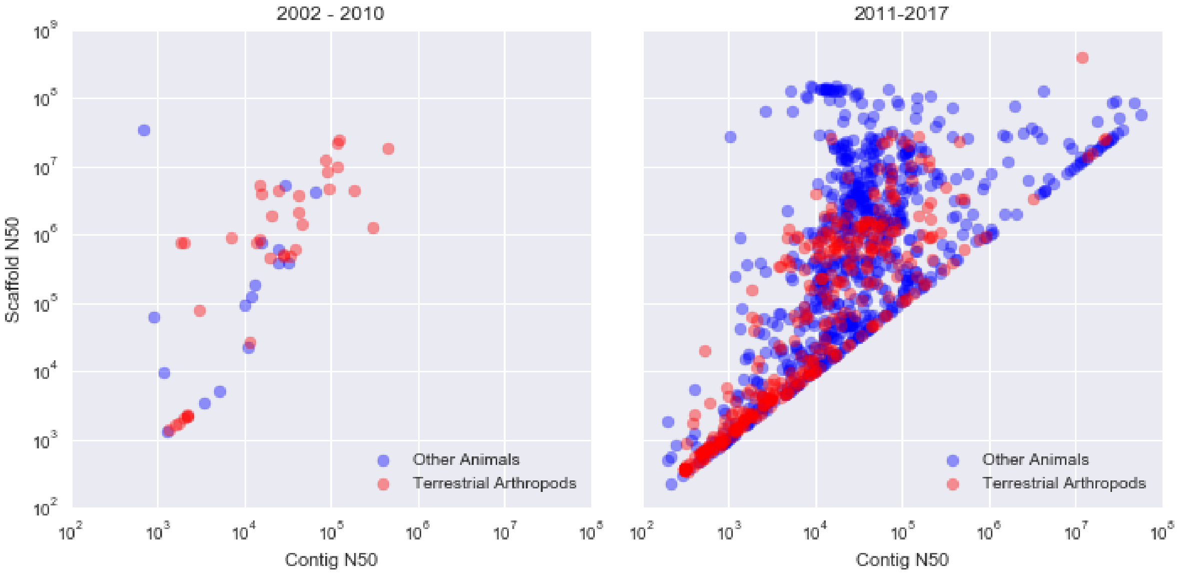
Blue:
Global Genome Biodiversity Network

Yellow:
National Museum of Natural History

Green:
Catalogue of Life

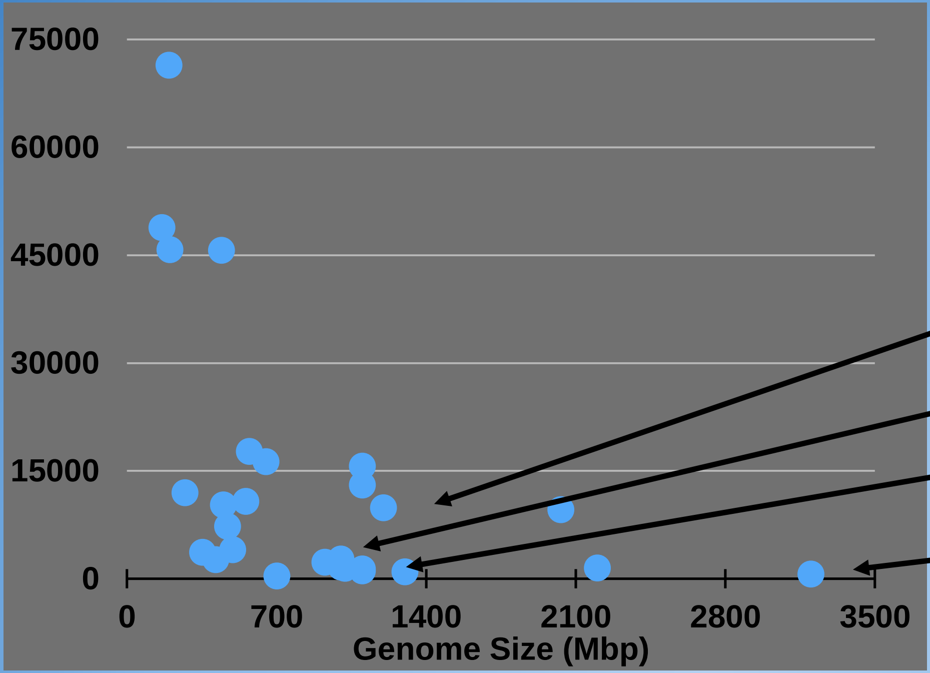
Terrestrial Arthropod Genomes





NG50 and Gene Number vs. Genome Size

NG50



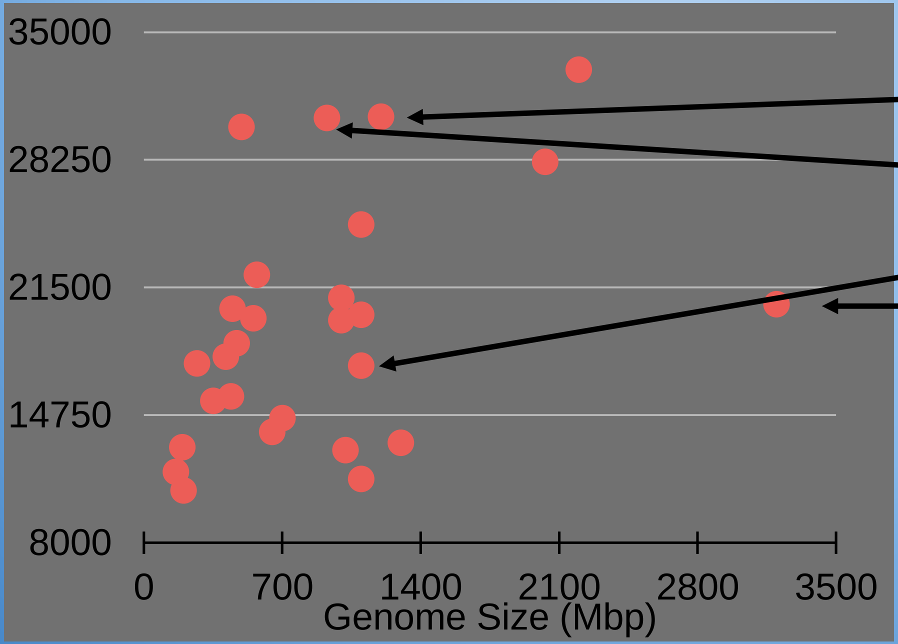
Common house spider

Bark scorpion

Black widow

Brown recluse

Gene Number



Common house spider

Bark scorpion

Black widow

Brown recluse

Dovetail Genomics: Genome Assembly

Estimated Dovetail physical coverage: 138X

	Starting Assembly	Final Assembly
Total Length	1443.9 Mb	1445.4 Mb
N50 Length	816 scaffolds; min 0.466 Mb	94 scaffolds; min 4.05 Mb
N90 Length	4824 scaffolds; min 0.025 Mb	448 scaffolds min 0.487 Mb

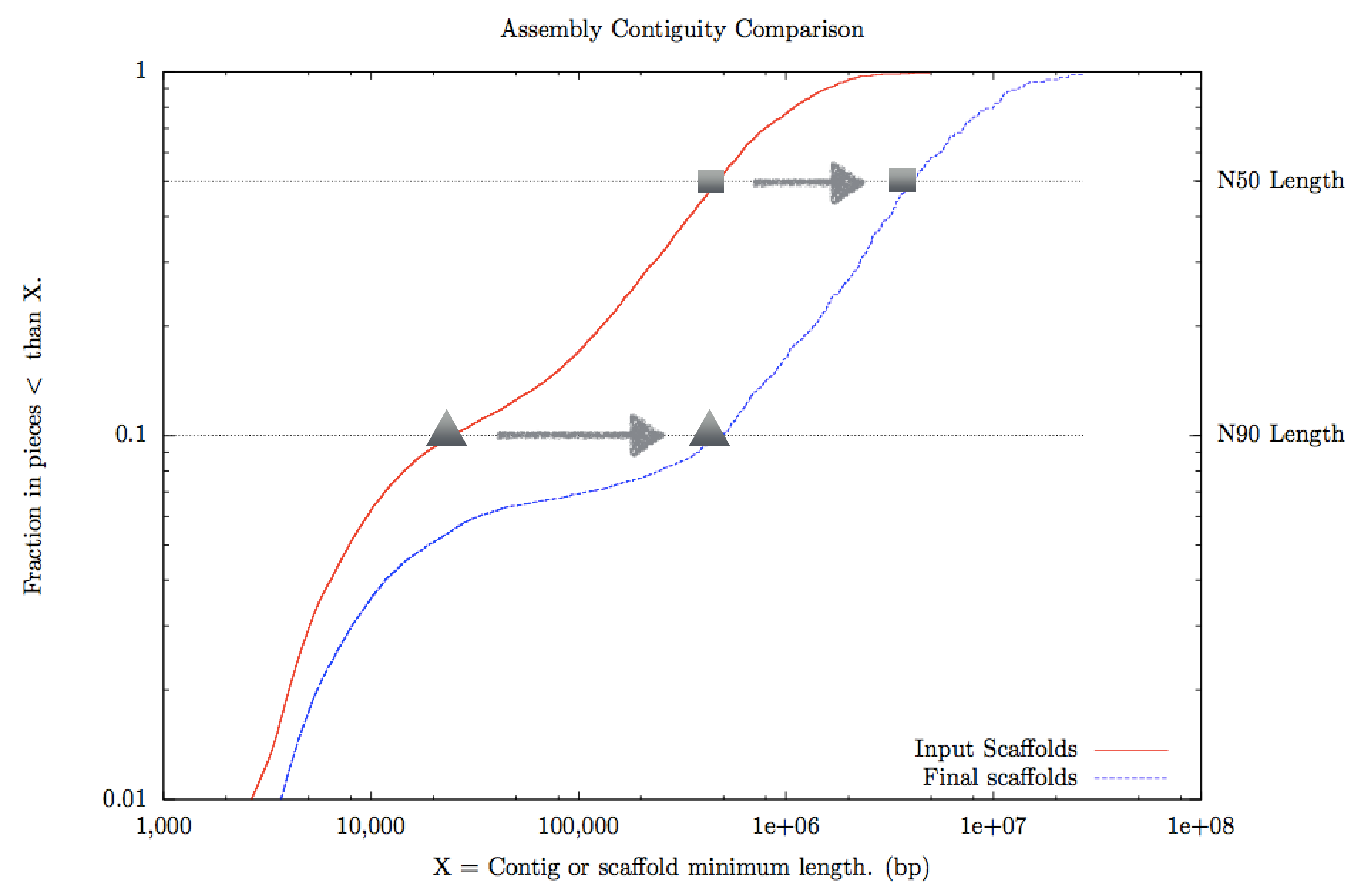
Ten-fold improvement

- Physical coverage: how many times on average is a given distance spanned by read pairs.
- N50: 50% of the genome is represented by N50 of scaffolds.
- N90: Same as above but 90% of the genome is represented by this number.

Testing new technologies: Dovetail Genomics

- Chicago pipeline obtains physical mapping data, involves *in vitro* chromatin assembly to condense DNA
- Constructs long-range sequencing libraries. Inserts span all distances up to DNA fragment size (library insert): 100 -150 kb for this library.

Dovetail Genomics: Scaffolding Contiguity



Signal density is lower in smaller scaffolds (i.e. there were **larger improvements in longer starting scaffolds**

Parasteatoda Synteny

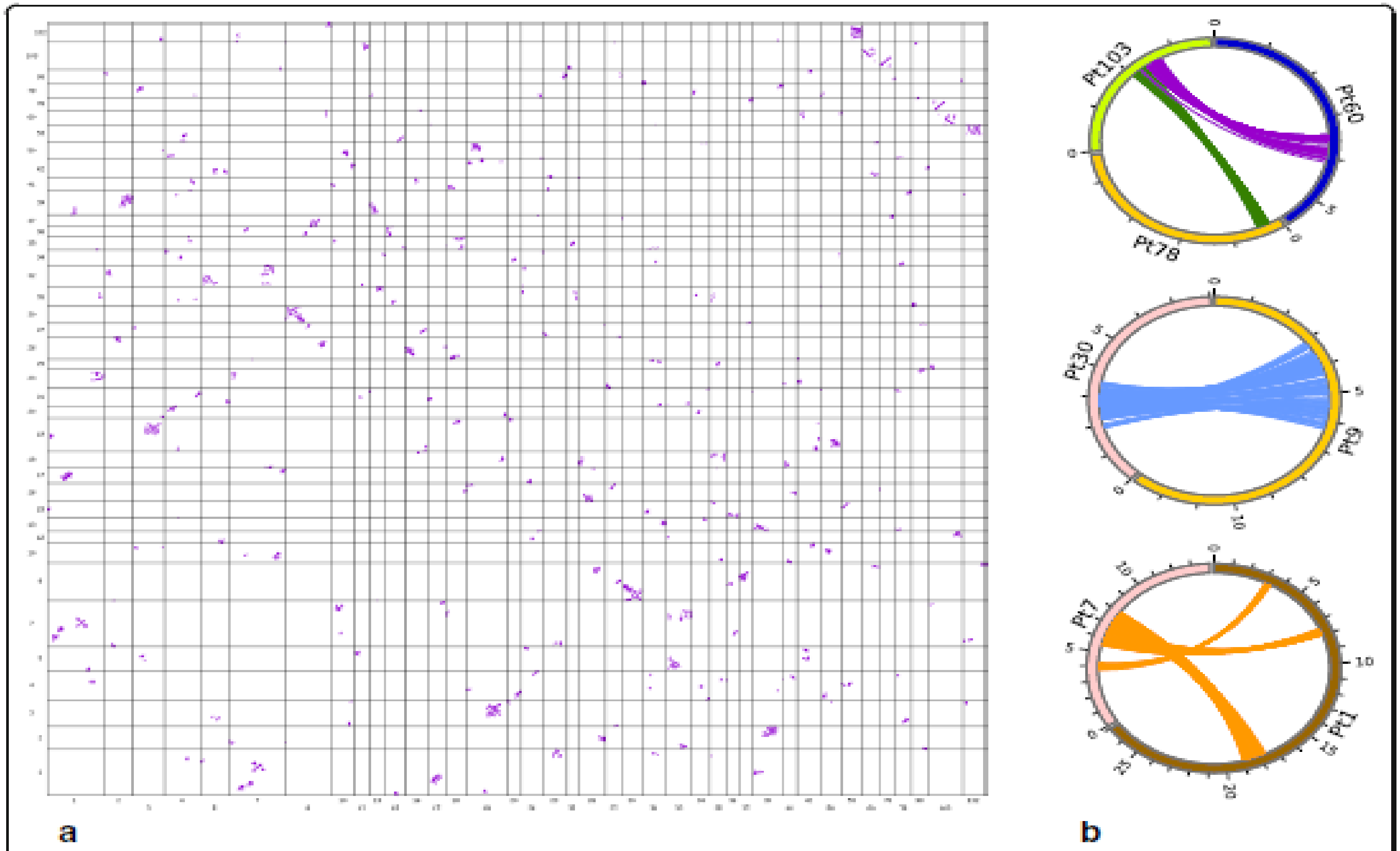


Fig. 5 Genome-scale conservation of synteny among *P. tepidariorum* scaffolds reveals signatures of an ancient WGD. **a** Oxford grid displaying the collinearity detected by SatsumaSynteny among the 39 scaffolds presenting the greatest numbers of hits on one another. On this grid (not drawn to scale), each point represents a pair of identical or nearly identical 4096-bp regions. Alignments of points reveal large segmental duplications suggestive of a whole-genome duplication event along with other rearrangements such as inversions, translocations and tandem duplications. **b** Circos close-ups of some of the colinearity relationships revealed by the Oxford grid

Schwager & Al. 2017. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution.

Nephila Stats

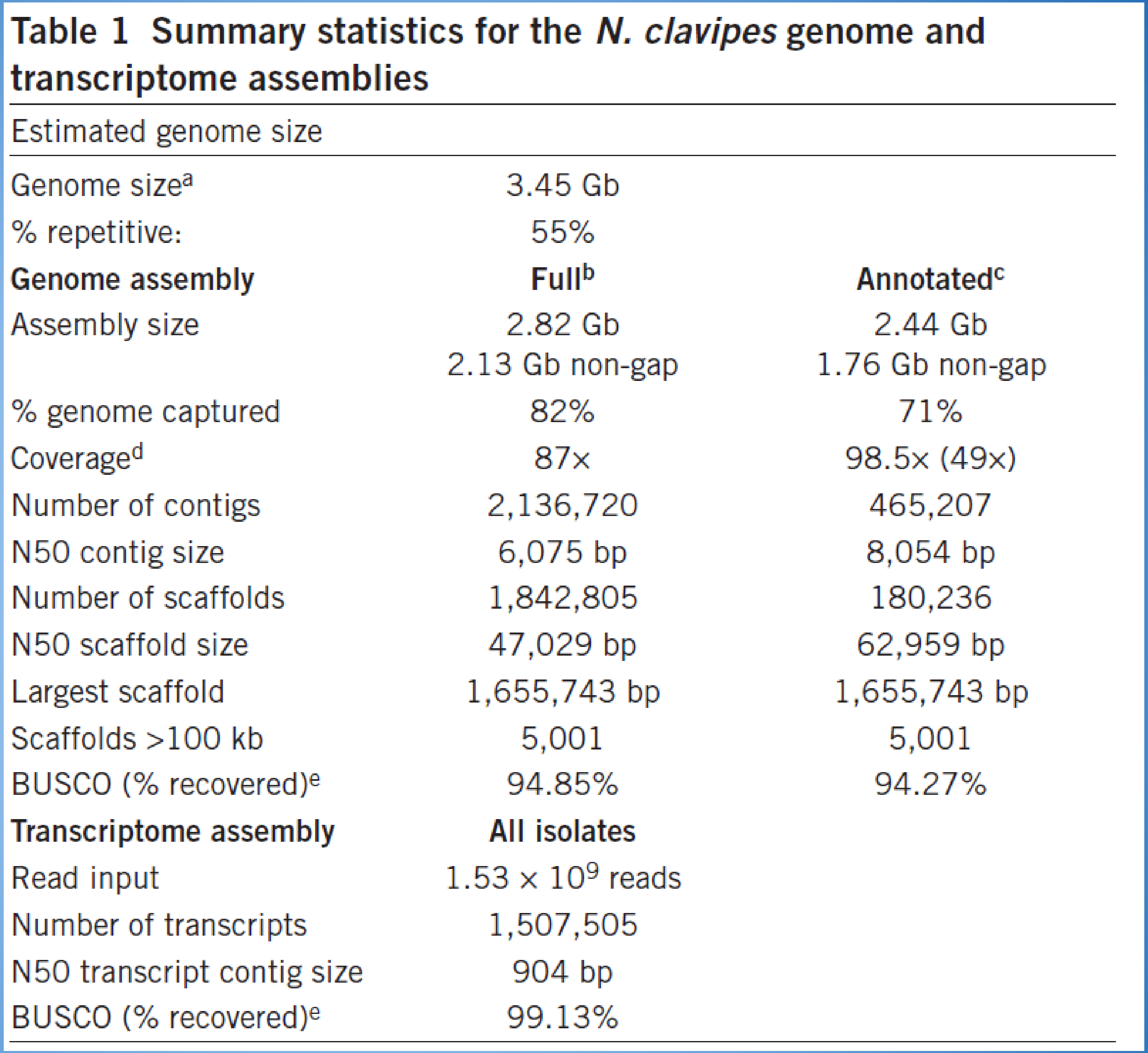
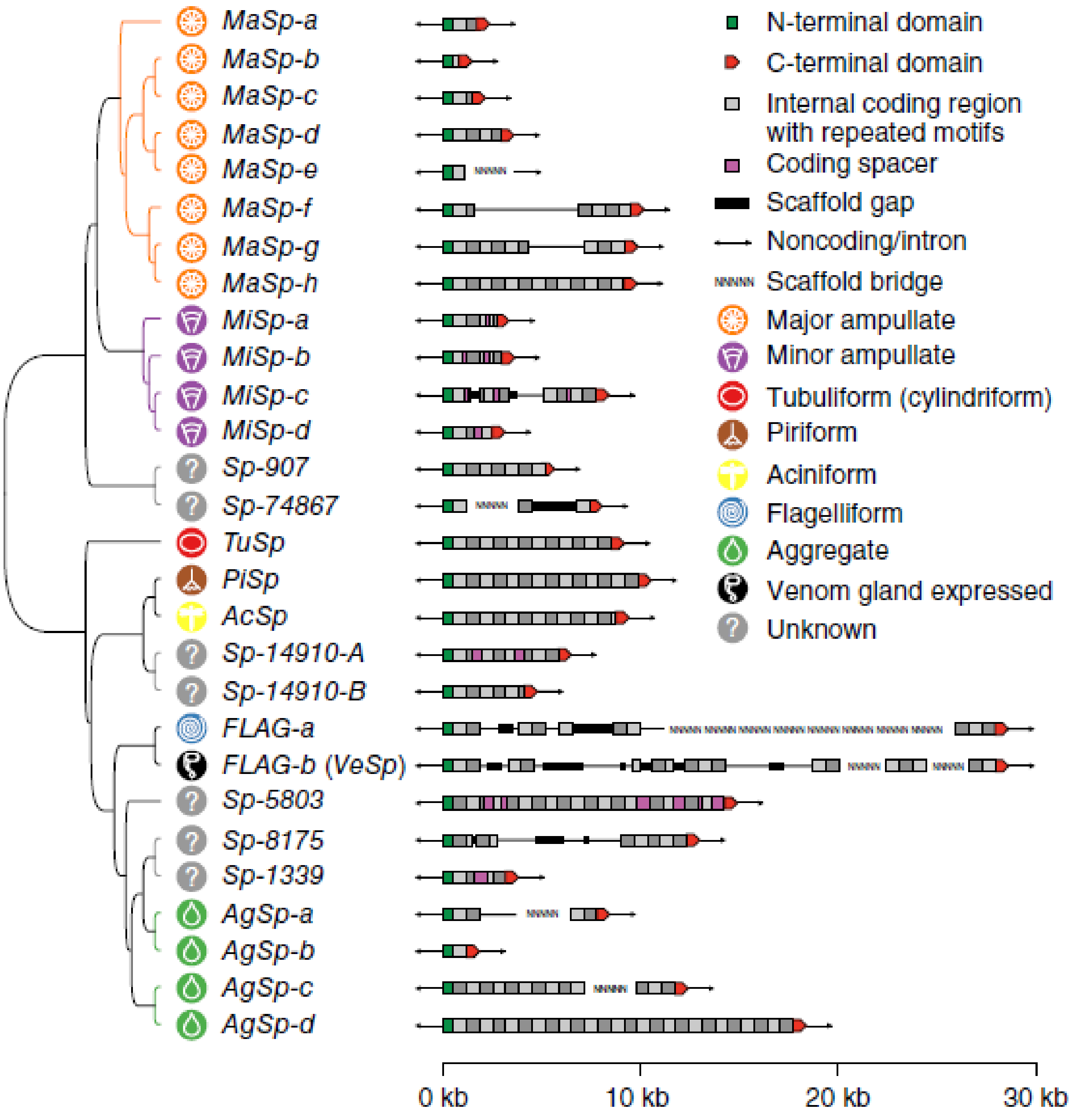


Table 1 Summary statistics for the *N. clavipes* genome and transcriptome assemblies

Estimated genome size		
Genome size[a]	3.45 Gb	
% repetitive:	55%	
Genome assembly	**Full[b]**	**Annotated[c]**
Assembly size	2.82 Gb 2.13 Gb non-gap	2.44 Gb 1.76 Gb non-gap
% genome captured	82%	71%
Coverage[d]	87×	98.5× (49×)
Number of contigs	2,136,720	465,207
N50 contig size	6,075 bp	8,054 bp
Number of scaffolds	1,842,805	180,236
N50 scaffold size	47,029 bp	62,959 bp
Largest scaffold	1,655,743 bp	1,655,743 bp
Scaffolds >100 kb	5,001	5,001
BUSCO (% recovered)[e]	94.85%	94.27%
Transcriptome assembly	**All isolates**	
Read input	1.53×10^9 reads	
Number of transcripts	1,507,505	
N50 transcript contig size	904 bp	
BUSCO (% recovered)[e]	99.13%	

Babb & Al. 2017. The Nephila clavipes genome highlights the diversity of spider silk genes and their complex expression

Nephila Silk Genes



Babb & Al. 2017. The Nephila clavipes genome highlights the diversity of spider silk genes and their complex expression

Thanks!

Seán Brady, Carol Butler, Katie Barker, Matt McDermott, Tom Orrell, Lee Weigt, Robert Costello, Chris Elias, Loretta Cooper, Bob Corrigan, Cyndy Parr, Chris Meyer, John Kress, Mike Ruggiero, GGBN partners….

Smithsonian
National Museum of Natural History

Association of American Universities 1900 AAU

Natural History Museum

Ocean Genome Legacy

iDigBio
Integrated Digitized Biocollections

neon

NCBI
National Center for Biotechnology Information

NSF

Open Tree of Life

ITIS Integrated Taxonomic Information System

Convention on Biological Diversity

Global Genome Biodiversity Network (GGBN)

Barcode of Life
Identifying Species with DNA Barcoding

eol
Encyclopedia of Life

Smithsonian
National Zoological Park

Consortium for the Barcode of Life

BIOCODE moorea
17 30 5 149 50 W

Smithsonian Tropical Research Institute
Center for Tropical Forest Science